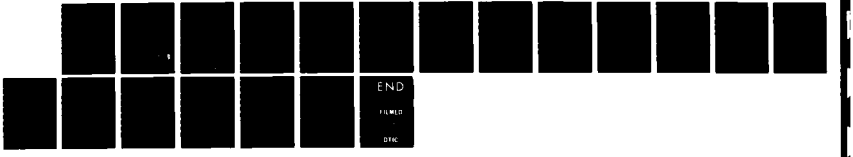
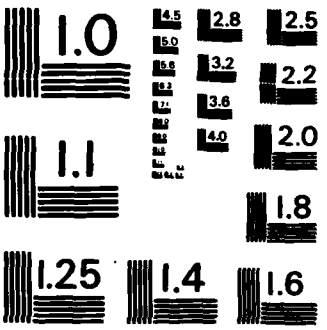
AD-A162 015 AN ALTERNATIVE PROCEDURE FOR ESTIMATING UNIT LEARNING CURVES(U) OFFICE OF THE COMPTROLLER (NAVY) WASHINGTON DC H W DAGEL SEP 85 1/1

UNCLASSIFIED F/G 5/1 NL





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

AN ALTERNATIVE PROCEDURE FOR ESTIMATING UNIT LEARNING CURVES

COST AND ECONOMIC ANALYSIS DIVISION
NCD-5

SEPTEMBER 1985
HAROLD W. DAGEL
OPERATIONS RESEARCH ANALYST

1. INTRODUCTION

The unit learning curve plays a prominent role in DOD cost analysis. In those cases when the model accurately describes the real-life situation, i.e., when the model is properly applied to the data, it can be a powerful tool for predicting unit production costs. There are, however, some unique estimation problems inherent in the model.

The usual method of generating predicted unit production costs attempts to extend properties of least squares estimators to non-linear functions of these estimators. The result is biased estimates of unit production costs. Another problem common to many learning curve applications is estimating lot midpoints and slope coefficients when both estimates depend on each other and both quantities are unknown.

This paper addresses the two problems discussed above and presents an alternative procedure for estimating unit learning curves. A simple modification to the usual estimators results in new estimators which yield unbiased estimates of unit production costs. The lot midpoint problem is overcome by another simple and widely used estimation technique, that of iterative ordinary least squares (OLS).

2. BACKGROUND

The learning curve equation is frequently employed in DoD cost analysis. Although there are several variations of the general form of the equation, the one considered here is that of a unit learning curve;

$$y = ax^b \qquad (2.1)$$

where y refers to the cost of unit x of a specified manufactured item and a and b are parameters to be estimated. Frequently a is referred to as the cost of unit one or the T_1 value, since when x = 1, y = a, regardless of the value of b. In learning curve applications b is a negative exponent usually ranging between zero and one in absolute value. Hence, as the number of units increases, the unit cost will decrease.

The stochastic[1] model corresponding to the functional model (2.1) is usually assumed to be

$$y = ax^b e^u \qquad (2.2)$$

---

1. The model (2.1) is a mathematical function while the stochastic model (2.2) includes the disturbance term. Stochastic in this case is meant to imply random.
2. The letter e is defined to be based of the natural logorithms, i.e., e = 2.71828 . . .

which includes a multiplicative function[2] of the disturbance term u. This error[3] term is assumed to be a well-behaved random variable with zero mean and constant variance σ^2. Thus, y has a lognormal distribution. (See Appendix A).

Given these results and a fixed value (but any fixed value) of the explanatory variable x, the mean or expected value[4] of y is

$$E(y) = ax^b e^{0.5\sigma^2} \qquad (2.3)$$

the median[5] of y is

$$M(y) = ax^b \qquad (2.4)$$

---

3. Disturbance term and error term are frequently used interchangeably and refer to the randomness in y not accounted for by the functional form of the equation.
4. The mean or expected value of y can be interpreted as the average value of y observed from repeated observation on the same value of x. It is usually denoted by the letter E.
5. The median of y is the "middle value" of y; or the value of y such that half of the observations are greater in value than y and half are less in value. It is usually denoted by the letter M.

Details of these derivations can be found in Goldberger [1][6]. Note that the mean and median of y are not the same. This is due to the fact that the lognormal distribution is not symmetric.[7] Recall that in the case of a linear regression function with additive error term the median is equal to the mean, provided the error term is normally distributed.

Since equation (2.2) is not linear in the parameters a and b, one cannot estimate these parameters by simple linear regression. The usual solution to this problem is to proceed as follows:

1) Transform equation (2.1) by taking logs[8] to yield

$$\ln y = \ln a + b \ln x \qquad (2.5)$$

---

6. A detailed and rigorous theoretical account of much of the underlying theory upon which this paper is based can be found in a paper by Arthur Goldberger which appeared in Econometrica in 1968.

7. A random variable which has a symmetric distribution has the same probability of being n units above the mean as it does of being n units below the mean. An example of a symmetric distribution is the normal distribution.

8. The letters ln are meant to represent the natural logorithm of the expression following. Natural logorithms are used exclusively throughout this paper.

2) Perform ordinary least squares on the linear equation (2.5) to obtain the estimated equation

$$\widehat{\ln y} = \widehat{\ln a} + \hat{b} \ln x \qquad (2.6)$$

where the "hats" denote OLS estimates.

3) Take the antilog[9] of the right-hand side of equation (2.6) and use the resulting equation to generate estimates of y in the original unlogged equation, viz.,

$$\hat{y} = e^{\hat{w}} \qquad (2.7)$$

where $\hat{w} = \widehat{\ln a} + \hat{b} \ln x$.

The least squares estimates obtained in step 2) are unbiased[10] and have the minimum variance of any unbiased estimators. These desirable properties result from the least squares estimation technique; see, for example, Draper and Smith [2, p.87]. It is important to note here that these properties apply

---

9. To take the antilog of a logorithm one merely raises e to the power of the logorithm. Hence, $e^{\ln x} = x$.

10. An estimator is said to be unbiased if its expected value is equal to the parameter it is meant to estimate. Thus, $E(x) = x$ implies that x is an unbiased estimator of x.

only to the parameters themselves and not to exponential functions of these parameters. Indeed, because of the convexity[11] of the exponential function,

$$E(e^{\hat{w}}) = e^{w + 0.5 \text{ var } (\hat{w})} \quad (2.8)$$

$$= a\, x^{b}\, e^{0.5 m^{*}\sigma^{2}} \quad (2.9)$$

where $m^{*}\sigma^{2}$ is the variance of a predicted y value for any given value of x. Specifically, this variance is the familiar covariance result,

$$m^{*}\sigma^{2} = \text{var } (\ln \hat{a}) + (\ln x)^{2} \text{ var } (\hat{b}) + 2 \ln x \text{ cov } (\ln \hat{a}, \hat{b}).$$

Thus, in step 3), $e^{\hat{\Delta}}$ is not an unbiased estimator of either the mean or the median of y. (Compare equation (2.9) with equations (2.3) and (2.4)).

---

11. The expected value of a convex function of a random variable is always greater than the convex function of the expected value of that random variable, i.e., if x is a random variable then $E(e^{x}) > e^{E(x)}$. See Mood, et al [3, p.72], for details.

3. UNBIASED ESTIMATORS

It is clear from the results of the last section that the customary procedure for estimating learning curves results in biased estimates. It would be nice if we could modify the estimator $e^{\hat{w}}$ discussed in the last section in such a manner so as to yield unbiased estimates of the mean and median of y. Fortunately, such a modification has been developed by Goldberger [1].

Goldberger has developed correction factors (See appendix B), F_M and F_E such that

$$E(F_M) = e^{-0.5m^*\sigma^2} \qquad (3.1)$$

$$E(F_E) = e^{0.5\sigma^2}\ e^{-0.5m^*\sigma^2} \qquad (3.2)$$

The products of the estimator $e^{\hat{w}}$ and the correction factors result in unbiased estimators of the mean and median of y, i.e.,

$$E(e^{\hat{w}}F_M) = e^{w}e^{0.5m^*\sigma^2}\ e^{-0.5m^*\sigma^2}$$

$$= e^{w} = ax^{b} \qquad (3.3)$$

$$E(e^{\hat{w}}F_E) = e^{w}e^{0.5m^*\sigma^2}\ e^{0.5\sigma^2}\ e^{-0.5m^*\sigma^2} \qquad (3.4)$$

$$= e^{w}e^{0.5\sigma^2}$$

$$= ax^b e^{0.5\sigma^2}$$

Comparing equations (3.3) and (3.4) with equations (2.3) and (2.4) will verify the fact that these estimators are unbiased.

Although the correction factors F_M and F_E involve infinite sums, they converge rather quickly to some preassigned tolerance. This property of rapid convergence and the use of digital computers make these estimators a desirable alternative to traditional estimation techniques.

4. LOT MIDPOINTS

Having solved the problem of biased estimators, we consider another problem associated with estimating unit learning curves. The unit learning curve relates unit cost or labor hours to the number of items produced. DOD contractors, however, usually account for costs by the lot rather than by the unit. Hence, while the average cost of a lot is known, the quantity associated with it, or the lot's midpoint, is not. Specifically, the midpoint of a lot will lie somewhere between the lot's first and arithmetic midpoint, with the exact location depending on the curve's slope. An unfortunate dilemma therefore emerges: lot midpoints can't be computed without knowing the slope, and the slope can't be estimated without knowing the lot midpoints.

In an attempt to resolve this perplexing problem, the following procedure is suggested:

1) Estimate the slope using approximations to true lot midpoints.

2) Compute new lot midpoints based on the previously estimated slope.

3) Repeat steps 1) and 2) until the delta between successive estimates of the slope is smaller than some preassigned tolerance.

In a sampling experiment, Flynn [5], examined the properties of the OLS estimator of a unit learning curve when lot midpoints are iteratively estimated.

The results of this sampling experiment showed that mean iterative OLS values were always very close to mean non-iterative values based on true lot midpoints. The iterative OLS estimator appeared to be unbiased in the samples examined. The estimator is not without problems, however, for sometimes it fails to converge. (48 out of 12,000 regression equations in the sampling experiment) Frequency of failure seems to increase as R^2 decreases and as lot quantities increase.

Fortunately, R^2 values in real-world learning curve estimation are typically very strong, thus minimizing the chance of non-convergence. In the rare event that iterative OLS breaks down, common rules of thumb can be applied to compute lot midpoints. It was also found that this estimator's iterative algorithm usually converges to four decimal places after only three to five iterations.

5. AN ALTERNATIVE ESTIMATION PROCEDURE

Based on the results of the preceding sections, the following estimation procedure is suggested as a desirable alternative to the customary procedure.

1) Transform equation (2.1) by taking logs to yield

$$\ln y = \ln a + b \ln x \qquad (5.1)$$

2) Use iterative OLS to estimate lot midpoint quantities and the OLS estimates

$$\widehat{\ln y} = \widehat{\ln a} + \hat{b} \ln x \qquad (5.2)$$
$$= \hat{w}$$

3) To predict new y values for any given x value use the estimators

$$\hat{y}_E = e^{\hat{w}} F_E$$

$$\hat{y}_M = e^{\hat{w}} F_M$$

depending on whether an estimate of the mean or median is wanted. The resulting estimates are unbiased and have performed consistently in sampling experiments.

To get an idea of the magnitude of the bias introduced by the customary estimation procedure, we illustrate a sample of 10 learning curve data sets. These data sets are chosen from Navy aircraft and missile acquisition programs. They were not chosen randomly, but instead, were chosen to indicate the wide range of bias possible using the usual estimation procedure.

Figure 5.1 shows estimated percent bias for median predicted T_1 values of the 10 samples when compared to the unbiased procedure suggested in this paper. Note how the estimated percent bias is directly related to the variance of the predicted y values, viz., $m^* \sigma^2$. In this case the predicted y values and T_1 values are the same, since x = 1.

Sample Size	Variance of e^W	Estimated % Bias
10	.00275	.14
8	.00429	.21
8	.00454	.28
8	.01053	.53
3	.01071	.54
10	.01204	.60
14	.08141	4.17
3	.11253	5.90
5	.30523	17.07
4	.72146	49.13

Figure 5.1 - Estimated % Bias of Median T_1 Values

Flynn [5] and Dagel [4] have developed computer software to implement the alternative estimation procedure described in this section. The FORTRAN-77 code is relatively compact and is currently being run on a VAX computer. It should present few problems to adapt this code to personal computers such as the IBM-PC.

APPENDIX A

A NOTE ON THE MEAN AND VARIANCE OF A LOGNORMALLY DISTRIBUTED RANDOM VARIABLE.

If Y is a positive random variable and if we define a new random variable

$$X = \ln Y$$

with X having a normal distribution, i.e. $X \sim N(\mu, \sigma^2)$ then

$$Y = e^X$$

has a lognormal distribution. The probability density function (p.d.f.) of X is

$$f_X(x) = (1/\sigma\sqrt{2\pi}) \exp[-(x-\mu)^2/2\sigma^2].$$

Hence the mean of Y can be derived in a straight - forward manner by evaluating the integral

$$E(y) = \int_{-\infty}^{+\infty} e^X f_X(x)\, dx \qquad [3, pp.176-177)$$

$$= e^{\mu + 0.5\sigma^2}$$

The variance of Y can be derived in a similiar manner by evaluating $E(Y^2)$ and then using the fact that

$$\text{var}(Y) = E(Y^2) - (E(Y))^2$$

This results in

$$\text{var}(Y) = e^{2\mu + 2\sigma^2} - e^{\sigma^2 + 2\mu}$$

Note that when $X \sim N(0, \sigma^2)$, i.e. when $\mu = 0$,

$$E(Y) = e^{0.5\sigma^2}$$

$$\text{var}(Y) = e^{2\sigma^2} - e^{\sigma^2}.$$

Details of these derivations can be found in Dagel [4].

APPENDIX B

The exact form of the correction factor as given by Goldberger is

$$F(w:v,c) = \sum_{j=0}^{\infty} f_j (cw)^j / j!$$

where

$$f_j = (v/2)^j \, \Gamma(v/2) \, / \, \Gamma[(v/2) + j \,].$$

and where

w = variance of the estimator

c = a constant

v = degrees of freedom.

The correction factor for the median of y is

$$F_M(w:v,c)$$

where

$$w = m^* s^2$$

$$c = -\tfrac{1}{2}.$$

The correction factor for the mean of y is

$$F_E\ (w{:}v,c)$$

where

$$w = 0.5s^2$$

$$c = (1-m^*)$$

Details of these derivations can be found in Goldberger [1].

REFERENCES

[1] Goldberger, Arthur S., "The Interpretation and Estimation of Cobb-Douglas Functions," Econometrica, Vol. 35, July-October, 1968, pp. 464-472.

[2] Draper, Norman and Smith, Harry, "Applied Regression Analysis," New York: John Wiley and Sons, Second Edition, 1981.

[3] Mood, A.M., Graybill, F.A. and Boes, D.C., "Introduction to the Theory of Statistics," New York: McGraw-Hill, Third Edition, 1974.

[4] Dagel, Harold W., "Unbiased Estimation of Power-Function Equations," Unpublished paper; Comptroller of the Navy, Cost and Economic Analysis Division, Washington, DC 20350.

[5] Flynn, Brian J., "Learning-Curve Properties when Lot Midpoints are Iteratively Estimated," Unpublished paper: Comptroller of the Navy, Cost and Economic Analysis Division, Washington, DC 20350.

END

FILMED

1-86

DTIC